Review
A Few Key Skills

• Design a good scoring function
  – Log likelihood ratio
  – Evaluate the significance (Bayesian or p-value)
• FDR
• Dynamic programming.
  – Sequence, tree, set, mass ...
• HMM.
• Fast searching
  – Suffix tree & array.
  – Filtration (Spaced seed)
• Working with NP-hardness
  – Approximation, Local search, Branch and bound
• Information distance.
Scoring Function

• Design a good scoring function
  – Log likelihood ratio
  – Estimate foreground and background probability (BLOSUM)
  – Relative entropy
  – Evaluate the significance (Bayesian, p-value, E-value)

• Result validation
  – Optimality doesn’t mean reality.
  – Should throw away garbage results.
  – FDR
Dynamic Programming

• Build optimal solution from the optimal solution of a smaller sub-problem.

• A partial order is needed on the concerned sub-problems.
  – Sequence (alignment, HMM)
  – Tree (ancestor reconstruction)
  – Set (spaced seed sensitivity)
  – Mass (de novo sequencing)
HMM

• The model:
  – Finite hidden states, finite symbols.
  – Transition/Emission probabilities.

• Input: A sequence of observation (emitted symbols).

• Output: The most likely path of hidden states.

• Gene prediction, speech recognition.
Fast Searching

• Suffix Tree and Array
  – Linear space, linear time construction.
  – Support many efficient string operations
    • Substring query
    • Longest common substring
    • Maximal repeats
    • Maximal unique match
  – Skew algorithm for suffix sorting.

• Filtration
  – Spaced seed.
Working with NP-hardness

• Approximation
• Local search
• Branch and bound
• Parameterized algorithm (not covered in this course)
Information Distance

• A universal distance
• But not computable.
  – “Too good to be true.”
• Provides a universal way to design practical distance based on a certain compression algorithm.